Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

&

Saw Swee Hock School of Public Health
National University of Singapore, Singapore

STATISTICAL METHODS FOR THE DETECTION AND ANALYSES OF STRUCTURAL VARIANTS IN THE HUMAN GENOME

Shu Mei, Teo

Karolinska Institutet

NUS
National University of Singapore

Stockholm 2012

Published by Karolinska Institutet. Printed by Universitetsservice AB.

ISBN 978-91-7457-865-2

For Mum and Dad.
With love.

ABSTRACT

Structural variations (SVs) are an important and abundant source of variation in the human genome, encompassing a greater proportion of the genome as compared to single nucleotide polymorphisms (SNPs). This thesis investigates different aspects of SV analysis, focusing on copy number variations (CNVs) and regions of homozygosity (ROHs). It is divided into four main studies, each focusing on a different set of aims.

In Study I, Identification *of recurrent regions of copy-number variation across multiple individuals*, we develop an algorithm and software to identify common CNV regions using individually segmented data. The identified common regions allow us to investigate population characteristics of CNVs, as well as to perform association studies.

In Study II, *Multi-platform segmentation for joint detection of copy number variants*, we develop an algorithm to identify CNVs using intensity data from more than one platform. The algorithm is useful when researchers have data from multiple platforms on the same individual.

In Study III, *Regions of homozygosity in three Southeast-Asian populations*, we identify ROHs in three Singapore populations, namely the Chinese, Malays and Indians. We characterize the regions and provide population summary statistics. We also investigate the relationship between the occurrence of ROHs and haplotype frequency, regional linkage disequilibrium (LD) and positive selection. The results show that frequency of occurrence of ROHs is positively associated with haplotype frequency and regional LD. The majority of regions detected for recent positive selection and regions with differential LD between populations overlap with the ROH loci. When we consider both the location of the ROHs and the allelic form of the ROHs, we are able to separate the populations by principal component analysis, demonstrating that ROHs contain information on population structure and the demographic history of a population.

Last but not least, in Study IV, *Statistical challenges associated with detecting copy number variants with next-generation sequencing technology*, we describe and discuss areas of potential biases in CNV detection for each of four commonly used methods. In particular, we focus on issues pertaining to (1) mappability, (2) GC-content bias, (3) quality-control measures of reads, and (4) difficulties in identifying duplications. To gain insights to some of the issues discussed, we download real data from the 1000 Genomes Project and analyze it in terms of depth of coverage (DOC). We show examples of how reads in repeated regions can affect CNV detection, demonstrate current GC correction algorithms, investigate sensitivity of DOC algorithm before and after quality-control of reads and discuss reasons for which duplications are harder to detect than deletions.

PREFACE

I first started dabbling with genetic data during my 4th year as a Statistics undergraduate in 2007. I was working on the Affymetrix 500K SNP array, one of the densest SNP microarrays at that time. Barely 5 years later, there are arrays with more than 5 million SNPs, not to mention Next-generation sequencing arrays that produce billions of reads in a single run. The technologies to study genetics have certainly evolved very rapidly, bringing with it new challenges in terms of statistical and bioinformatics analyses.

When I first learnt of the term 'CNV', the concept sounded simple to me: That we have regions of the genome that are deleted/duplicated, and that based on the intensity of our measurements, less intense means less of that particular region, and vice versa. "Not too complex!" I thought naively. As I continue to learn more, the multitude of problems/challenges that comes associated with the analysis of noise-rich CNV data is enormous. As put across aptly by John Ioannidis on genetic data from microarrays in general, "…this noise is so data-rich that minimum, subtle, and unconscious manipulation can generate spurious "significant" biological findings that withstand validations by the best scientists, in the best journals. Biomedical science would then be entrenched in some ultramodern middle ages, where tons of noise is accepted as "knowledge". – The Lancet 365: 454-455.

Nevertheless, I hope that with these four years of hard work, I have helped made a little more sense out of the massive amount of genetic data we have.

LIST OF PUBLICATIONS

This thesis is based on the following original articles which will be referred to in the text by their Roman numerals.

I. **Teo SM**, Salim A, Calza S, Ku CS, Chia KS, Pawitan Y. (2010) Identification of recurrent regions of copy number variation across multiple individuals. *BMC Bioinformatics* **11**:147.

II. **Teo SM**, Pawitan Y, Kumar V, Thalamuthu A, Seielstad M, Chia KS, Salim A. (2011) Multi-platform Segmentation for joint detection of copy number variants. *Bioinformatics* **27**:11.

III. **Teo SM**, Ku CS, Salim A, Naidoo N, Chia KS, Pawitan Y. (2012) Regions of homozygosity in three Southeast Asian populations. *Journal of Human Genetics* **57**: 101-108.

IV. **Teo SM**, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variants with next-generation sequencing technology. *Manuscript.*

Other relevant publications:

- **Teo SM**, Ku CS, Naidoo N, Hall P, Chia KS, Salim A, Pawitan Y. (2011) A population-based study of copy number variants and regions of homozygosity in healthy Swedish individuals. *Journal of Human Genetics* **56**:524-533.

- Ku CS, **Teo SM**, Naidoo N, Sim X, Teo YY, Pawitan Y, Seielstad M, Chia KS, Salim A. (2011) Copy number polymorphisms in new HapMap III and Singapore populations. *Journal of Human Genetics* **56**:552-560.

- Ku CS, Naidoo N, **Teo SM**, Pawitan Y. (2011) Regions of homozygosity and their impact on complex diseases and traits. *Human Genetics* **129**:1-15.

- Ku CS, Naidoo N, **Teo SM**, Pawitan Y. (2011) Characterising Structural Variation by Means of Next-Generation Sequencing. *Encyclopedia of Life Sciences (ELS)*. John Wiley & Sons, Ltd: Chichester.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure 5.1: Diagram illustrating the non-triviality of determining if two CNVs are the 'same' variant. In (a), CNV1 and CNV2 overlap completely. In this case, we are confident that the two CNVs are the same. In (b), the start and end positions of CNV1 and CNV2 differs, but there is substantial overlap between the two. In (c), CNV1 is completely within the range of CNV2 but the two CNVs differ vastly in lengths. In most research papers, scientists are comfortable with using a 50% reciprocal overlap to determine if two CNVs are concordant.

LIST OF ABBREVIATIONS

The following abbreviations have been used in this thesis and in the associated four original publications:

aCGH	Array comparative genomic hybridization
AIC	Akaike information criterion
ANOVA	Analysis of variance
AS	Assembly based
BAF	B allele frequency
Bp	Base-pairs
CAHRES	Cancer Hormone Replacement Epidemiology in Sweden
CBS	Circular Binary Segmentation
COMPOSITE	Cumulative Overlap Using Very Reliable Regions
COVER	Cumulative Composite Confidence Scores
CNV	Copy number variation
CNVR	Copy number variation region
DGV	Database of Genomic Variants
DNA	Deoxyribonucleic acid
DOC	Depth of coverage
EHH	Extended haplotype homozygosity
FDR	False discovery rate
GWAS	Genome-wide association studies
HIV	Human immunodeficiency virus
HMM	Hidden Markov model
HTS	High throughput sequencing
HWE	Hardy Weinberg equilibrium
iHS	Integrated haplotype score
kb	Kilo base-pairs
LD	Linkage disequilibrium
LINEs	Long interspersed nuclear elements
LOH	Loss of homozygosity
LRR	Log R ratio

LTR	Long terminal repeat
MAF	Minor allele frequency
MPSS	Multi-platform smooth segmentation
MPCBS	Multiple platform circular binary segmentation
NGS	Next-generation sequencing
PEM	Paired end mapping
PCA	Principal component analysis
PCR	Polymerase chain reaction
QC	Quality-control
RD	Read depth
rDNA	Ribosomal deoxyribonucleic acid
RP	Read pair
ROH	Regions of homozygosity
SINEs	Short interspersed nuclear elements
SOLiD	Supported Oligonucletide Ligation Detection System
SMS	Single molecule sequencing
SNP	Single-nucleotide polymorphism
SR	Split read
SV	Structural variants
TGS	Third generation sequencing
VKORC1	Vitamin K epoxide reductase complex subunit 1
VNTR	Variable number of tandem repeats
WTCCC	Wellcome Trust Case Control Consortium

Chapter 1 – INTRODUCTION

Genetic variation in the human genome can take many forms, including single-nucleotide polymorphisms (SNPs), copy number variations (CNVs), indels, regions of homozygosity (ROHs), and other structural variants (SVs). In the last couple of years, genome-wide association studies (GWAS) have been widely used to correlate genetic differences to phenotypic variation, but they were largely focused on SNPs.

CNVs and other SVs were less appreciated until two landmark studies in 2004 identified widespread deletions and duplications in the human genome (Sebat *et al*., 2004; Iafrate *et al*., 2004). By now, CNVs are widely recognized as a prevalent form of variation in the genome, encompassing a greater proportion of the genome as compared to SNPs. An estimated 1.2% of a single genome differs from the reference human genome when considering CNVs, as compared to 0.1% by SNPs (Pang *et al*., 2010). Recent studies have found CNVs to be associated with complex diseases such as human immunodeficiency virus (HIV) infection, cancer, diabetes, mental disorders, obesity, Parkinson's disease and autoimmune diseases (Wain *et al*., 2009; The Wellcome Trust Case Control Consortium 2010). ROHs are also more abundant than previously thought (Gibson *et al*., 2006), and are associated with complex diseases such as schizophrenia and late-onset Alzheimer's disease (Lencz *et al*., 2007; Nalls *et al.*, 2009).

That, as compared to SNPs, the association of CNVs and ROHs with complex diseases is not as well-studied is in part due to greater complexity in identifying these multi-base, multi-allelic variants, and also greater complexity in performing

association studies with these variants. Early works on CNVs/ROHs have focused largely on identifying and characterizing regions in the genome which harbour them. This has been necessary in laying the foundation to improve our understanding of CNVs/ROHs for subsequent association analysis with human complex diseases.

The most common technologies for CNVs identification in the last couple of years are high density SNP arrays and array comparative genomic hybridization (aCGH) arrays; the former (SNP arrays) are also commonly used for detection of ROHs. However, the data generated from these techniques are noisy, and identifying CNVs comprehensively with high resolution still remains a technical and statistical challenge. aCGH and SNP arrays are also limited by the resolution of the array to determine precise locations of CNV breakpoints, and are unable to locate copy-neutral events such as inversions and translocations.

Sanger sequencing, often seen as the gold standard for CNV detection, is able to detect CNVs with higher accuracy and resolution, to detect balanced rearrangements such as inversions and translocations, as well as to detect CNVs in regions where probe density of other platforms is low. However, the technique is not feasible for a large number of genomes due to time and budget constraints. Next-generation sequencing (NGS) attempts to combine the benefits of array technology and sequencing. The biggest advantage of NGS over traditional Sanger sequencing is the ability to sequence millions of reads in a single run at a comparatively inexpensive cost (Metzker, 2010). However, with billions of reads generated per individual, there is an increasing need for more bioinformatics support and computers with larger storage and higher computing powers, and for such support to keep pace with the

rapidly changing technologies. Already, there is a great demand for information technology infrastructure and bioinformatics team to analyse the massive amount of data, with speculations that the costs associated with down-handling, storing and analysis of the data could be more than the production of the data.

There is still a need for the development of new statistical/bioinformatics methods and software for the systematic analysis of CNV/SV data. This is the focus of this thesis.

Chapter 2 – BACKGROUND

In this chapter, I will introduce some concepts in CNV/ROH analysis, including definitions and introduction to existing technology, software and algorithms in detection of CNV/ROH. These will facilitate the understanding of subsequent chapters.

2.1 Terminology and nomenclature

Human genetic variations refer to differences in the deoxyribonucleic acid (DNA) sequences among different individuals; they can take many forms, including single-nucleotide polymorphisms (SNPs), indels, copy number variations (CNVs), and other copy-neutral variations such as inversions, translocations and regions of homozygosity (ROHs). These genetic variations span a spectrum of sizes, ranging from 1 base-pair (bp) changes to whole chromosomal changes (e.g. aneuploidy). The occurrences of these genetic variations are attributed to different diverse mechanisms. For example, the predominant mechanisms for CNV formation include non-allelic homologous recombination and non-homologous end joining (Hastings *et al.,* 2009; Conrad *et al.,* 2010). ROHs are thought to be a result of autozygosity or uniparental isodisomy (Gibson *et al.,* 2006).

Table 2.1 summarizes the definitions of variants from single base changes to the sub-microscopic level (larger variants are not discussed). Note that the definitions for the different classes of genetic variants based on size are often unclear at the edges of each class. For example, larger indels may sometimes be termed CNVs even when their sizes are less than 1 kb.

Types of variation	Size	Definition*	Remarks
SNVs, SNPs, single-nucleotide insertions-deletions (indels)	1 bp	SNVs are variations of a single nucleotide (see Figure 1). When the variation is common (usually defined as having a frequency of more than 1%), we call it a SNP (Figure 2.1).	Most SNPs are single nucleotide substitutions, although single nucleotide deletions/insertions may also fall under this category.
Indels, microsatellites, minisatellites, inversions, di-,tri-tetranucleatide repeats, variable number of tandem repeats (VNTRs)	2 to < 1000 bp	Indels are typically defined as insertions or deletions that are smaller than 1 kb and larger than 1 bp.	The size cut off is rather arbitrary; Database of Genomic Variants (DGV) defines indels in the size range of 100 bp to 1 kb.
CNVs, segmental duplications, inversions, translocations	1000 bp to sub-microscopic	CNVs are additions or deletions in the number of copies of a segment of DNA (larger than 1 kb in length) when compared to a reference genome (Figure 2.2).	Some large indels larger than 500 bp may also be termed CNVs. Common CNV larger than 1% population frequency are termed copy number polymorphism (CNP).
ROHs	> 500 bp	ROHs are continuous stretches of the genome (usually more than 500 kb) without heterozygosity in the diploid state.	

Table 2.1: Definition of the different classes of genetic variations, partly adapted from Figure 1 of Scherer *et al.,* 2007. *only selected types of variation are defined.

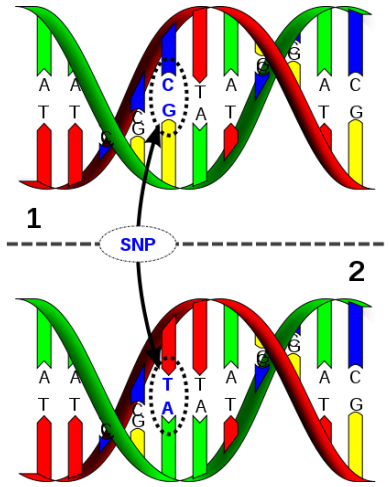

Figure 2.1: C-T single nucleotide variation.Source: http://en.wikipedia.org/wiki/File:Dna-SNP.svg

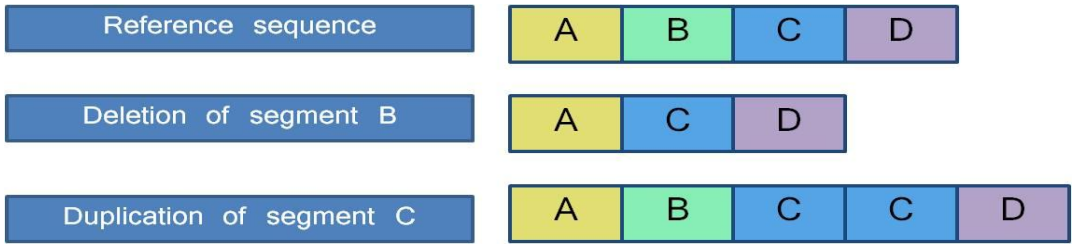

Figure 2.2: Schematic and simplified diagram of a deletion and duplication (adapted from Ku *et al*., 2010).

ROHs are sometimes termed loss of homozygosity (LOH), which includes hemizygous deletions (where there is only one copy of the region). Genotypes of SNPs within hemizygous deletions may be erroneously called as homozygous resulting in a region that may seem to be a ROH based on SNP genotypes alone. Figure 2.3 illustrates the differences in intensity patterns for ROH and one-copy deletion; while both ROH and one-copy deletion have similar B allele frequency (BAF) patterns, the Log R ratio (LRR) for ROH is around zero while it is below zero

for one-copy deletion. In this thesis, ROH always refer to the copy-neutral variant, where the region is in diploid state and all bases within the region are homozygous.

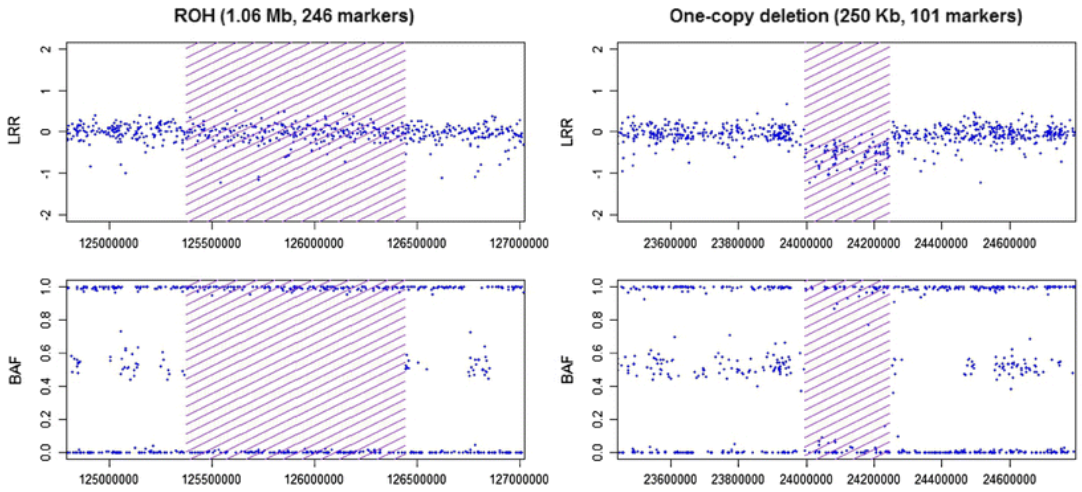

Figure 2.3: (Left panel) ROH signature with LRR around zero and no clusters at BAF of 0.5. (Right panel) One copy deletion signature with decreased LRR and similar pattern of BAF as ROH. The *x*-axis is the genomic probe location and each point represents a probe in the SNP array. (Figure from Ku *et al.*, 2011).

2.2 CNV and ROH detection technologies

In the last decade or so, the most commonly used technologies for CNV detection are whole-genome array comparative genome hybridization (aCGH) and high-density SNP arrays. ROHs are typically detected using high-density SNP arrays. CNVs/ROHs detected using these technologies are unfortunately limited by the density of the probes, as well as the location of the probes. For example, array platforms with more than 1 million probes have a lower detection limit of 10-25 kb in the size of CNV (McCarroll *et al*., 2008). Sanger sequencing provides better resolution and accuracy, but it is not cost/time-effective to use on a genome-wide scale for many individuals. The recent development of next generation sequencing (NGS) platforms that allow massive parallel sequencing have the potential to discover

smaller CNVs that were not previously discovered, detect balanced rearrangements such as inversions and translocations, as well as detect rare CNVs for which SNP arrays have no probes for. The biggest advantage over traditional Sanger sequencing is the ability to produce large amount of sequencing data in a single run.

However, as compared to SNPs, detection of CNVs is more challenging because of its complexity as a multi-base, multi-allelic variant. As a result, different algorithms and methods often give vastly different estimates in the number and breakpoints of CNVs. Currently, in the Database of Genomic Variants (DGV), there are more than 130,000 (merged) CNVs from 37 different studies, encompassing more than 52% of the genome; a likely gross overestimation of the true percentage of the genome encompassed by CNVs. This is because all the different studies use a heterogonous array of technologies, algorithms, filtering parameters, and samples.

2.3 CNV and ROH detection algorithms

Detection of CNVs from aCGH arrays is mostly based on locating change-points in intensity-ratio patterns that would partition each chromosome into several discrete segments. On the other hand, the hidden Markov model (HMM) is particularly popular for detection of CNVs from SNP arrays, where the hidden states provide a natural way of combining information from the total signal intensity (known as log R ratio, LRR) and the relative allele frequency (known as B allele frequency, BAF) values. Briefly, the HMM assumes several possible hidden states such as ‘deletion’, ‘normal’, ‘region of homozygosity’ and ‘duplication’ and analyse the most possible state-transition path, assuming that the copy numbers of nearby SNPs are dependent

(Wang *et al*., 2007). Illustrated in Figure 2.4, a 'normal copy' has three BAF clusters and the LRR is centred around zero; a ROH has LRR centred around zero but only two clusters at both extremes of the BAF.

The output from a CNV detection algorithm provides the following information: (1) Chromosome number (2) Start location (3) End location (4) Copy number. For example, this is a typical output from PennCNV:

chr6:32565228-32593190 numsnp=30 "length=27,963" "state1,cn=0"

It tells us that in Chromosome 6 of this individual, from the position 32565228 to position 32593190, there is a deletion where this individual has zero copies as compared to the reference panel. There are 30 probes in this region in the platform used, and the length of the region is 27,963 bases.

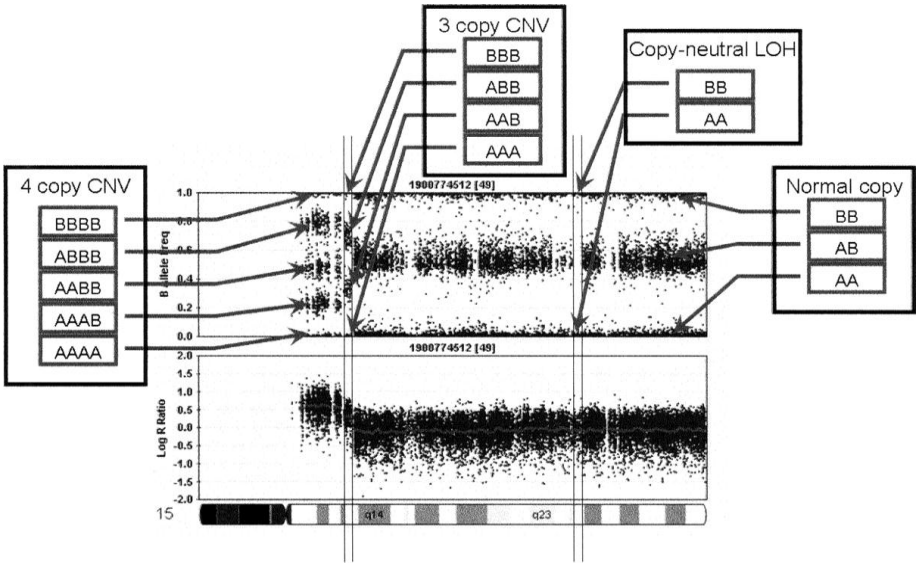

Figure 2.4: Figure from Wang *et al.*, 2007, illustrating the unique patterns in LRR and BAF of the different copy number states. A 'normal copy' has three BAF clusters and the LRR is centred around zero; a ROH has LRR centred around zero but only two clusters at both extremes of the BAF.

2.4 Sequencing technologies

2.4.1 First generation sequencing

First generation sequencing is typically referred to as 'Sanger sequencing', and is introduced by Frederick Sanger in 1977 (Sanger, 1977). It is the main form of sequencing technique used over the last 30 years until the arrival of next-generation sequencers in 2005. Sanger sequencing is able to sequence reads of length ~ 800-1000 bases (Hert *et al.*, 2008; Schloss *et al.*, 2008; Venter *et al*., 2001).

However, Sanger sequencing is laborious and costly; its inability to process more than 96 sequence reads at a time limits its application to large scale genome-wide sequencing efforts for many individuals (Mardis, 2008). For example, it took nearly ten years and three billion dollars to sequence the first human genome in the Human Genome Project (Schadt *et al.*, 2010).

2.4.2 Next-generation sequencing (NGS)

Next-generation sequencing (NGS) or also known as high-throughput sequencing (HTS) is able to simultaneously sequence millions of DNA reads. This ability to produce large amount of sequencing data in a single run at a comparatively inexpensive cost is its biggest advantage over traditional Sanger sequencing (Metzker, 2010). Currently available NGS sequencers in the market include the Roche 454 Genome Sequencer FLX System, Illumina Genome Analyzer, Illumina HiSeq and Applied Biosystems' Supported Oligonucletide Ligation Detection System (SOLiD).

NGS has the potential to discover smaller CNVs that were not previously discovered, to detect balanced rearrangements such as inversions and translocations, as well as to detect CNVs in regions where probe density of other platforms, such as SNP arrays, is low. NGS technologies have facilitated and accelerated the process of identifying genetic variations through whole-genome re-sequencing projects, including the 1000 Genomes Project.

However, there are some technical features of NGS that result in several challenges. Firstly, due to an effect called 'dephasing', there is an increase in noise and sequencing errors as the read length extends, thereby limiting the read lengths of NGS to ~35 – 400 bases (Schadt *et al.,* 2010). The short read lengths in turn complicate alignment and assembly. Secondly, in order to generate a large number of DNA molecules, polymerase chain reaction (PCR) amplification is required. This amplification process biases the frequency in which different portions of the genome are sequenced (Schadt *et al.,* 2010).

2.4.3 CNV detection using NGS

Broadly, there are four complementary methods for CNV detection using NGS data, namely (1) depth of coverage (DOC, also known as read-depth (RD) methods), (2) paired-end mapping (PEM), (3) split-read (SR) and (4) assembly-based (AS) methods (Alkan *et al.,* 2011). Except for the latter, the other three classes of methods require first mapping the sequenced reads to a known reference genome. The different methods are usually complementary to one another as the underlying concepts excel

at detecting certain types of variants, and a large proportion of discovered variants remain unique to a particular approach (Alkan *et al*., 2011).

Some algorithms use a combination of methods for more accurate detection of CNVs. For example, CNVer supplements DOC with PEM information in a unified framework (Medvedev *et al.,* 2010). Genome STRiP combines information from DOC, PEM, SR as well as other features of sequence data at population level (Handsaker *et al.,* 2011). Genome STRiP is one of the highest performing method used in the 1000 Genomes pilot Project, indicating that there is benefit in combining different approaches (Mills *et al*., 2011).

Depth of coverage

DOC methods typically count the number of reads that fall in each pre-specified window of a certain size (Abyzov *et al*., 2011; Yoon *et al*., 2009). The underlying concept of identifying CNVs using DOC is similar is that of using intensity data: a lower than expected DOC /intensity indicates deletion and a higher than expected DOC /intensity indicates duplication (Figure 2.5). The algorithm relies heavily on the assumption that the sequencing process is uniform, i.e., the number of reads mapping to a region is proportional to the number of copies. However, certain biases such as GC-content and mappability cause this assumption to be unrealistic; regions of the genome may be over or under-sampled regardless of the copy number of the region, often resulting in spurious signals. DOC algorithms usually detect large CNVs and are unable to detect copy neutral events such as inversions and translocations. Single-end or paired-end data may be used for this analysis.

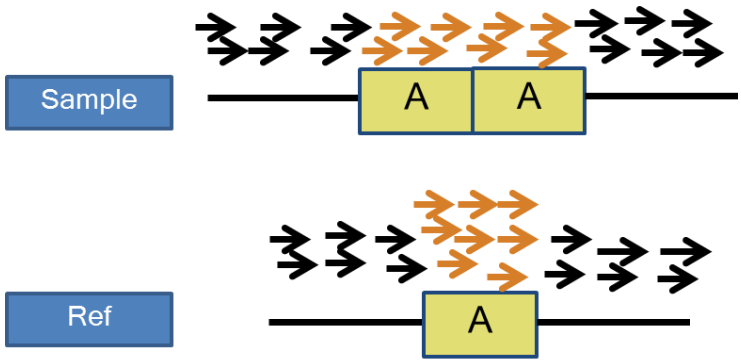

Figure 2.5: Schematic diagram illustrating the concept of depth of coverage method for CNV detection. If the sample has an additional copy relative to the reference genome, when the reads are mapped to the reference, we would observe an increase in depth of coverage in the region.

Paired-end mapping

PEM methods require the reads to be paired (Chen *et al.*, 2009). The concept is that the fragments of DNA from which the reads are to be sequenced have a fragment length (or also known as insert size) of a certain distribution, and a longer than expected fragment length indicates a deletion in the studied genome compared to a reference while a shorter than expected fragment length indicates an insertion. Based on the patterns from which the paired reads are mapped to the reference, read pair analysis can also detect inversions and translocations. The size of CNVs detected using PEM is limited by the insert size and as a result, PEM often detects smaller CNVs. For example, PEM does not allow the discovery of insertions larger than the insert size (Dalca *et al.*, 2010).

Split-read

SR methods uses paired reads as well. They focus on pairs of reads where one read is mapped to the reference while the other read failed to be aligned (Ye *et al.*, 2009).

The idea is that where the location of the unmapped read may span the breakpoint of the CNV. SR analysis has the advantage of being able to pinpoint the location of the breakpoints.

Assembly-based

AS methods, on the other hand, do not align the reads to a known reference but construct the genome piece-by-piece, which is known as *de novo* sequencing. Some AS methods use the reference genome as a guide to resolve repeats. This is known as *comparative assembly* (Pop *et al.,* 2004). AS methods can discover new non-reference sequence insertions. AS methods works best for small genomes such as bacterial genomes and are less widely used in NGS sequencing of humans because the short reads from NGS makes assembly in repeat regions difficult (Ye *et al*., 2009). Even though assembly algorithms continue to improve, due to the short read lengths, de novo sequencing using NGS are still not capable of achieving similar quality as that using Sanger sequencing (Schadt *et al*., 2010).

2.5 Repetitive DNA

Repetitive DNA refers to sequences that are highly similar or identical to sequences in other parts of the genome. They are abundant in the human genome and covers almost 50% of the human genome (Treangen *et al.,* 2012). Table 2.2 summarises repeat type, number, percentage of genome covered and approximate length of each repeat class. The repeat type is broadly characterized into tandem or interspersed repeats where the former refer to repeats that are adjacent to each other while the latter refers to repeats that are separated by hundreds, thousands or millions of bases.

In next generation sequencing, reads from repetitive regions may map equally well to several locations in the reference genome. Due to the ambiguity in the alignment step, these reads often cause problems in SNP and SV detection. Reads that can be mapped equally well to more than one location are termed *multi-reads*.

Repeat class	Repeat type	Number	% genome	Length (bp)
Minisatellite, microsatellite or satellite	Tandem	426, 918	3%	2 -100
SINEs	Interspersed	1, 797, 575	15%	100 - 300
DNA transposon	Interspersed	463, 776	3%	200 - 2000
LTR retrotransposon	Interspersed	718, 125	9%	200 - 5000
LINEs	Interspersed	1, 506, 845	21%	500 - 8000
rDNA	Tandem	698	0.01%	2000 - 43000
Segmental duplications and other classes	Tandem or interspersed	2, 270	0.20%	1000 - 100000

Table 2.2: This table summarises for each repeat class, the repeat type (tandem or interspersed), number in the hg19 human genome, percentage of the hg19 human genome covered, and approximate lower and upper bounds for the lengths of the repeat. (Table adapted from Treangen *et al.,* 2012). Short interspersed nuclear elements (SINEs), Long terminal repeat (LTR), Long interspersed nuclear elements (LINEs), ribosomal DNA (rDNA).

2.6 Copy number variation region (CNVR)

CNVR or also known as CNV loci or common CNV or recurrent CNV are CNVs that occur in the same/similar location across several individuals. Most CNV detection algorithms identify CNVs individual-by-individual, but common CNVs are known to exist among different individuals. However, the identification of the individual-specific CNVs is not precise, especially in terms of the breakpoints. This poses a challenge when we want to summarize the population characteristics or perform association studies, because it is unclear if CNV1 from individual 1 describes biologically the same event as CNV2 from individual 2 if their breakpoints do not match exactly.

2.7 Hardy Weinberg Equilibrium of CNVR

Suppose a bi-allelic SNP has allele frequencies p and q (where $p+q = 1$) for alleles a and b respectively, regardless of gender. Assuming random mating, in the next generation, the frequencies of the aa, ab and bb genotypes are p^2, $2pq$ and q^2 respectively. The allele frequencies of a and b have not changed and remain p and q, such that in the following generation, the genotype frequencies will again be p^2, $2pq$ and q^2, and so forth. This is known as Hardy Weinberg Equilibrium (HWE); i.e., that the frequency of alleles and genotypes remain constant from generation to generation in a large population assuming random mating. The Pearson's chi-squared test is typically used to test for departure from these expected frequencies, indicating violation of HWE.

Since it has been observed that the majority of common CNV regions are inherited (Locke *et al.*, 2006), we expect, for a population of normal, healthy individuals, the integer copy numbers for the majority of CNVRs to be in HWE. This is supported by McCarroll *et al*., (2008)'s study that found that 98% of common bi-allelic CNVRs do not violate HWE. McCarroll *et al*., (2008) also found that about 90% of common CNVs are bi-allelic.

In principal, HWE applies to both bi-allelic CNVRs and multi-allelic CNVRs. Bi-allelic CNVs are those with only two alleles, forming three possible copy numbers. For example, CNVs with copy numbers 0, 1, 2 or 2, 3, 4 are considered bi-allelic. Multi-allelic CNVs are those with more than two alleles, for example, with alleles '0', '1' and '2', the possible copy numbers are 0, 1, 2, 3 and 4. Testing HWE for bi-allelic

CNVs is straightforward and similar to the test for SNPs. However, for multi-allelic CNVRs, HWE test cannot be performed directly on the unphased copy-number because there is an issue with different combinations of alleles producing the same copy-number. For example, with alleles '0', '1' and '2', the copy number 2 can have genotype (1, 1) or (0, 2).

2.8 GWAS of CNVs

Genome-wide association studies (GWAS) using SNPs have been widely performed over the last couple of years, resulting in over 1400 published associations (at $p \leq 5x10^{-8}$) for 237 traits (from the National Human Genome Research Institute: http://www.genome.gov/26525384). This is in part due to greater accuracy and completeness with which SNPs, as compared to CNVs, can be assayed.

Earlier studies on CNV discovery have paved the way for subsequent association studies of CNVs. For example, the Wellcome Trust Case Control Consortium (WTCCC) performed a large scale GWAS study of CNVs in 16000 cases of eight common diseases using a customized aCGH that was designed based on previously identified CNVs (Wellcome Trust Case Control Consortium, 2010). The WTCCC study found several CNV loci to be associated with Crohn's disease, rheumatoid arthritis, type 1 diabetes and type 2 diabetes. However, these loci have been previously identified through SNP based GWAS, reflecting the observation that common CNVs are well tagged by SNPs.

2.9 Linkage disequilibrium

The non-random association of alleles at two or more loci in the genome is known as linkage disequilibrium (LD), i.e. that the occurrences of some combinations of alleles at two or more loci are more or less frequent than expected based on their individual allele frequencies. For example, suppose allele A_1 at SNP A and allele B_1 at SNP B have frequencies p_1 and q_1 respectively. If the two SNPs are independent, then we expect to see the A_1B_1 haplotype at a frequency of p_1q_1; any departure from this expected frequency means that the two SNPs are in LD. Most commonly used statistics to quantify the extent of LD between two loci are the r^2 and D' statistics (Lewontin *et al*., 1960). Both statistics are based on the extent of departure of the observed haplotype frequency from the expected. Let x_{11} be the observed AB haplotype frequency. Then, $D = x_{11} - p_1q_1$. Now, let the other two alleles of SNP A and SNP B have frequencies p_2 and q_2 respectively.

$$r^2 = \frac{D^2}{p_1p_2q_1q_2} \text{ and } D' = \frac{D}{D_{max}} \text{ where } D_{max} = \begin{cases} \min(p_1q_1, p_2q_2) & \text{when D} < 0 \\ \min(p_1q_2, p_2q_1) & \text{when D} > 0 \end{cases}$$

Both measures have a minimum value of 0, which indicates independence between the two loci, and maximum value of 1, which indicates complete dependence between the two loci.

2.10 Quantification of positive selection

Positive selection is the phenomenon where certain variants rise to a frequency at a faster rate than would be expected, i.e., the favouring of variants that increase survival and reproduction. Under neutral evolution, new variants need a long time to reach

high frequency, resulting in common variants usually having short range LD because recombination would have occurred to disrupt the haplotypes (Sabeti *et al.,* 2002). Hence, one 'clue' or signature that provides evidence of positive selection is an unusually long and common haplotype which indicates an allele which rose to high frequency rapidly before recombination occurs (Bersaglieri *et al.*, 2004).

One statistic used to quantify positive selection is the integrated haplotype score (iHS) (Voight *et al.*, 2006). Briefly, this score measures how unusual the haplotypes around a core SNP are, relative to the rest of the genome. The iHS first utilizes the extended haplotype homozygosity (EHH) statistic (Sabeti *et al.,* 2002); the EHH measures the decay of haplotype identity as a function of distance. For each SNP, haplotype homozygosity starts at 1 and decays to zero with increasing distance. Alleles under selection tend to have high haplotype homozygosity that extends much further, resulting in a large area under the EHH curve. The iHS is a standardized measure of the integrated EHH. Clusters of SNPs with large positive or large negative iHS are evidence of position selection in the region.

Chapter 3 – AIMS

Overall, the general aim of this thesis is to use and develop statistical and bioinformatics methods to improve detection and analyses of structural variants. The thesis is divided into four studies as follows:

I. We develop a method and accompanying software to identify common CNV regions in multiple individuals. The identified common regions can be used for downstream analyses such as group comparisons in association studies.

II. We develop a method and software to identify CNVs by using data from multiple platforms simultaneously. We also propose an objective criterion for discrete segmentation required for downstream analyses. For each identified segment, the software reports a p-value to indicate the likelihood of the segment being a true CNV.

III. We investigate the population characteristics of ROHs in three Singapore populations (Chinese, Malays and Indians), and access the relationship between the occurrence of ROHs and haplotype frequency, regional LD and positive selection.

IV. We highlight problems and issues encountered when analysing NGS data for CNVs, in particular, those pertaining to DOC methods. We use real data from the 1000 Genomes Project to highlight and investigate challenges associated with (1) GC-content, (2) quality score of reads, and (3) identifying CNVs in repeated regions.

Chapter 4 - PAPER SUMMARIES

4.1 Study I: Identification of recurrent regions of copy-number variation across multiple individuals.

4.1.1 Motivation

Most algorithms for CNV-detection detect CNVs sample-by-sample with individual specific breakpoints. However, common CNV regions (CNVRs) are likely to occur at the same genomic locations across multiple individuals.

4.1.2 Methods overview

The main novelty of our algorithm is that we exploited the region specific confidence score statistic provided by commonly used segmentation programs, PennCNV and QuantiSNP. This statistic indicates how likely the detected CNV for a particular individual is true. By not incorporating the use of individual specific confidence scores, it means that all regions contribute equally to the statistic used to identify the common regions, but some regions are more likely to be true positives than others. Our method utilizes both the confidence score statistic, as well as the frequency of occurrence, to identify CNVRs. Intuitively, we have less confidence in a CNV that occurs in one individual than one that occurs in many individuals. However, a single occurrence of CNV might still be a true discovery if it is associated with a high confidence score, i.e., it is based on a strong signal. Since individual CNVs span different probes, the number of individual regions that overlap each probe varies. However, common CNV regions tend to occur at almost the same genomic locations

across multiple individuals. Hence, we expect the common regions to be identified by consecutive probes where a 'significant' number of individuals have an overlapping CNV. Furthermore, we also expect the confidence score of the individual regions to be relatively high.

Method 1: Cumulative Overlap Using Very Reliable Regions (COVER)

To calculate the COVER statistic for a probe, we sum the number of high-confidence individual regions that overlap that probe. The common region is then defined as consecutive probes for which the COVER statistic is greater than or equal to a specified threshold, *u*. Users provide two parameters here: the confidence score threshold, *c*, to determine high-confidence regions and *u*, the threshold for the COVER statistic.

Method 2: Cumulative Composite Confidence Scores (COMPOSITE)

In COVER, we may miss regions that are detected with lower confidence scores but nonetheless detected consistently across a large number of individuals. For the COMPOSITE statistic, we sum all individual regions that overlap the probe, weighted by their confidence score.

Method 3: Clustering of Individual CNV regions within a Common Region

The CLUSTER method uses a clustering algorithm that further refines the regions identified by either method 1 or method 2. This method is motivated by the observation of a complex mixture of sub-regions within a CNVR identified by COVER/COMPOSITE (Figure 4.1).

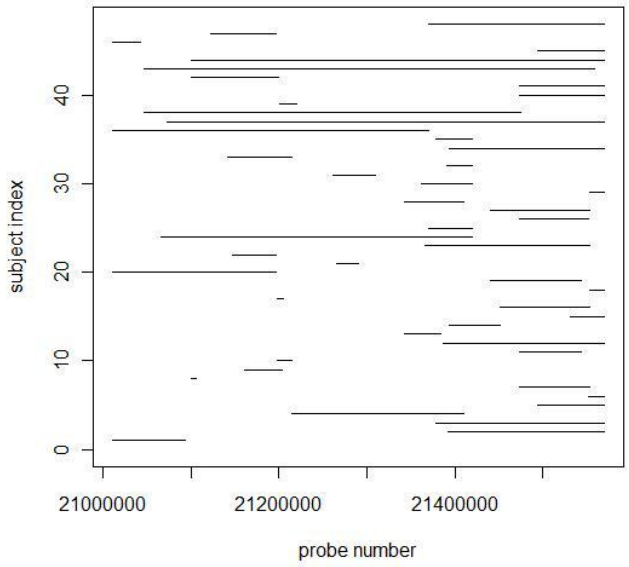

Figure 4.1: An example of a CNVR identified by COVER. We observe that despite being identified as a common region, the individual regions still portray a mixture phenomenon of several distinct sub-regions (from Teo *et al*., 2010).

4.1.3 Results

Comparison with sequenced regions

To assess the performance of our methods, we use 112 HapMap samples and vary the threshold parameters in our methods. For each threshold, we calculate discordance rates with sequencing-based results (Kidd *et al.,* 2008) and rates of departure from HWE. The discordance rates as well as the rates of departure from HWE decrease when we select CNVs with higher confidence scores, showing the importance of further processing of the CNVs (for COVER results, see Figure 4.2). Similar results were observed for COMPOSITE method (Figure not shown). Concordance rates improve after refinement with CLUSTER.

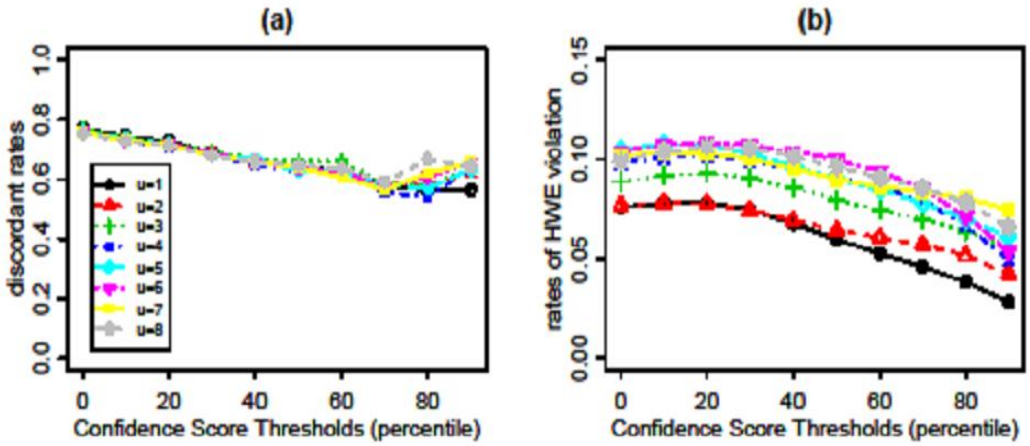

Figure 4.2: (a) Discordance rates for COVER method decreases as the confidence score thresholds increase. (b) Rates of departure from HWE decreases as the confidence score thresholds increase (from Teo *et al*., 2010).

Comparison to other algorithms

We compare our methods to two previously published methods, STAC (Diskin *et al*., 2006) and GISTIC (Beroukhim *et al.,* 2007). We find that our methods are better at identifying low-frequency but high-confidence CNV regions.

Implementation

The methods are implemented in an R package, *cnvpack*. The main input is a list of detected individual CNV regions with the following information: Sample name, chromosome number, detected integer copy number, start and end genomic locations and a confidence score. The package can be downloaded from http://www.meb.ki.se/~yudpaw.

4.2 Study II: Multi-platform segmentation for joint detection of copy number variants.

4.2.1 Motivation

At the time this research was carried out, SNP genotyping platforms from major commercial companies, such as Illumina and Affymetrix, were rapidly evolving, and it is not uncommon for research groups to have data from different platforms for the same individuals. For CNV detection, marker density is one important factor. Different platforms have different sets of marker panels and combining data from multiple platforms would undoubtedly give higher marker density. It has the potential to yield more precise and accurate detection of CNVs and its breakpoints. However, combining such data is not straightforward as different platforms show different degrees of attenuation of the true copy-number, noise characteristics and marker panels (Zhang *et al.*, 2010). There is still a relative lack of formal procedures for combining information from different platforms for copy-number calling. Most studies with data from multiple platforms interrogating the same samples usually process the data independently for each platform, after which the identified segments are combined in an ad-hoc manner. This approach does not fully utilize information from the different platforms, and when the segmented results from the different platforms differ, it is difficult for researchers to come to a consensus in a statistically rigorous manner.

In this study, we develop a new method for identifying CNVs by using data from multiple platforms simultaneously. As we are often interested in discrete segments of

CNVs for downstream analyses, we also develop an objective method to obtain discrete segments, and provide a p-value associated with each segment; the p-value would indicate how likely the segment is a CNV, and can be used to filter false positives.

4.2.2 Methods overview

The method, multi-platform smooth segmentation (MPSS) is an extension of Huang *et al*. (2007)'s single-platform *smoothseg* algorithm which is based on the Cauchy random-effect model that allows jumps in the underlying copy-number patterns. MPSS uses normalized $\log_2$-intensity ratios from two or more platforms and estimates the underlying copy number pattern for an individual. For each individual, we denote $\{x_1, \ldots, x_n\}$ as the union of the probe locations from the different platforms, with $x_1 < x_2 < \ldots < x_n$. Denote $\{y_{x1j}, \ldots, y_{xnj}\}$ as the set of $\log_2$-intensity ratios from platform j. We write our model as

$$y_{xij} = f(x_{ij}) + e_{xij} ,$$

where f is a random effects parameter that is common to all platforms, meaning that each platform is assumed to measure the same underlying copy-number pattern; as such, background normalization is recommended so that data from the different platforms become comparable. The error term e_{xij} is platform-specific to take into account different noise characteristics of the different platforms. The platform specific error structure was chosen to be *t*-distributed to incorporate a heavy tailed structure that can deal with outliers in the observations. The smoothness of f can be

expressed by assuming that the scaled second order differences $a_i^* \equiv \frac{\Delta^2 f_t}{(\Delta x_t)^2}$ are independent and identically distributed with some distribution. We specify a_i^* to follow the Cauchy distribution to allow for jumps in the segments. To estimate the random-effects parameter *f*, we derive an iterative weighted least squares algorithm by maximizing the likelihood of the Cauchy random-effects model.

4.2.3 Results

We compare *MPSS* against the single-platform *smoothseg* algorithm, an existing multiplatform method, called *MPCBS* (Zhang *et al*., 2010), and its associated single platform method, *CBS* (Olshen *et al*., 2004). We use nine HapMap samples, which were previously genotyped by both the Illumina 1M and Affymetrix 6.0 SNP arrays by our collaborators at the Genome Institute of Singapore, Agency for Science, Technology and Research. For the same samples, we have the integer copy-numbers from Conrad *et al.* (2010)'s study, which we use as a reference list.

When signals from the different platforms are consistent, we get increased power to detect the CNVs when we combine information from different platforms, especially in areas where a single platform has low density of probes (Figure 4.3a) or complete lack of probes (Figure 4.3b). To compare against other methods, we perform individual-specific comparisons with the reference list and report the number of overlapping bases as a proportion of the total length of CNVs identified by the method and as a proportion of the total length of CNVs in the reference list. In Figure 4.4, we show that MPSS CNVs have greater amount of overlap with the reference, indicating better performance.

Implementation

The algorithm is implemented in an R package *MPSS* that can be freely downloaded from http://www.meb.ki.se/~yudpaw. The main inputs are vectors of genomic positions, chromosome numbers and $\log_2$-intensity ratios from each platform.

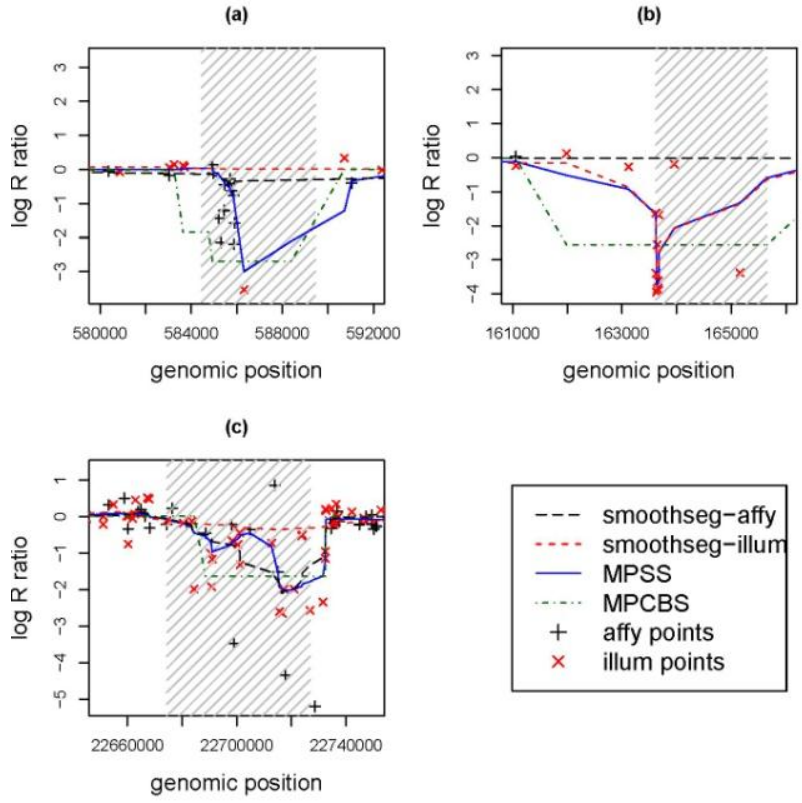

Figure 4.3: Examples of segments detected by the multiplatform methods. (a) A deletion in Chromosome 8. Single platform smoothseg on Illumina platform was unable to identify the deletion due to lack of probes in the region. Single platform smoothseg on Affymetrix platform was unable to identify the deletion due to insufficient signal. (b) A deletion in Chromosome 16. Single platform smoothseg on Affymetrix platform was unable to identify the deletion due to complete lack of probes in the region. (c) A deletion in Chromosome 22 (from Teo *et al.,* 2011).

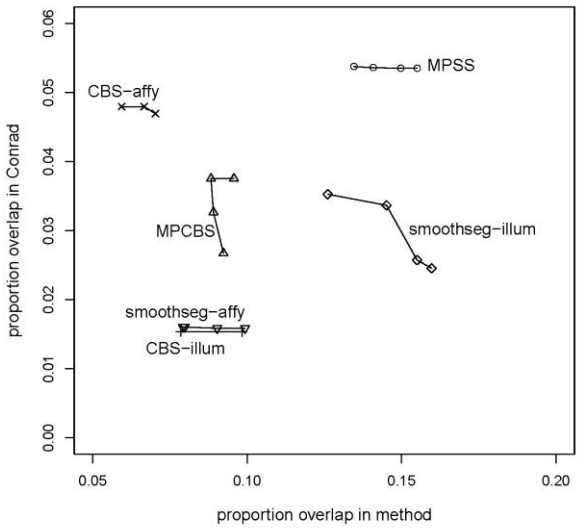

Figure 4.4: The number of overlapping bases as a proportion of Conrad's CNVs and as a proportion of each method's CNVs; the different points for each method correspond to the different thresholds. A higher proportion of overlap indicates better performance (from Teo *et al.,* 2011).

4.3 Study III: Regions of homozygosity (ROHs) in three Southeast Asian populations

4.3.1 Motivation

The genomes of outbred populations were first shown in 2006 to contain an abundance of long stretches > 500kb without heterozygosity (Gibson *et al.,* 2006; Li *et al.,* 2006). Since then, there have been several studies that investigate the population characteristics of ROHs in healthy individuals (McQuillan *et al.,* 2008; Nothnagel *et al.,* 2010; O'Dushlaine *et al.,* 2010), and also several studies that perform association analyses to identify ROHs that are associated with complex diseases (Yang *et al.,* 2010; Lencz *et al.,* 2007; Nalls *et al.,* 2009). However, the majority of these studies are conducted on European populations, and there is a lack of knowledge of ROHs in Asian populations. Thus, the first aim is to characterize ROHs in the three main Singapore populations, namely the Chinese, Malays and Indians.

Furthermore, it was observed that the location of ROHs is markedly non-random, where unrelated individuals may share similar region boundaries. Some loci are caused by a single common haplotypes, whereas others are a consequence of several common haplotypes that could be markedly disparate (Curtis *et al.,* 2008). The second aim of this study is to investigate the relationship between the occurrence of ROHs and haplotype frequency, linkage disequilibrium (LD) and positive selection.

4.3.2 Samples

The genomic DNA samples used in this study were part of the Singapore Genome Variation Project[1], whose aim was to characterize the extent of common genetic polymorphisms and the haplotypes in each of the three ethnic groups in Singapore (Teo *et al.,* 2009). Peripheral blood DNA was extracted from a total of 292 individuals and genotyped using the Illumina Human 1M Beadchip and the Affymetrix Genome-wide Human SNP Array 6.0.

4.3.3 Results

We identified an average of 207, 179 and 126 ROHs per individual for Chinese, Malays and Indians respectively. Indians have lower numbers as well as lower total length of ROHs as compared to Chinese and Malays. About 83% of the ROHs are within the 500 kb to 1 Mb size range while 17% of them are greater than 1 Mb.

Using the individual regions to form common regions (using the software from Study I), we obtain 1256 common ROH loci in the three populations. We study the relationship of the common ROHs with haplotype frequency, LD and positive selection. For each locus, we test for differences among the 3 populations in terms of ROH frequencies and haplotype frequencies, and 47 loci (<4%) differ significantly in frequencies while 899 loci (69%) differ significantly in haplotype frequencies among the populations. One interesting example is a 700 kb region in Chromosome 16 that overlaps with the Vitamin K epoxide reductase complex subunit 1 (VKORC1) gene,

[1] approved by the National University of Singapore – Institutional review Board (Reference Code: 07 – 199E)

where genetic polymorphisms within this gene has been found to correlate with differences in warfarin dosage and response (Aquilante *et al*., 2006; Harrington *et al*., 2005). In the Singapore populations, the Indians were observed to display warfarin resistance, thus requiring a higher dose as compared to the Chinese and Malays (Zhu *et al.,* 2007). The ROH frequencies of this region are 21%, 13% and 20% for the Chinese, Malays and Indians respectively (no significant difference in frequencies). However, the haplotypes frequencies of this region among the three populations differ drastically (Table 4.1), especially between the Indians and the other two populations.

	Haplotype A	Haplotype B
Chinese	0.31	0.0052
Malay	0.28	0.045
Indian	0.0060	0.34

Table 4.1: Haplotype frequencies of three populations in an ROH that overlaps VKORC1 gene (from Teo *et al.,* 2012).

With regards to haplotype frequency and regional LD, we find that the frequency of an ROH is positively associated with the total frequency of the top three haplotypes as well as with regional LD. The majority of regions detected for recent positive selection and regions with differential LD between populations overlap with the ROH loci. When we consider both the location of the ROHs and the allelic form of the ROHs, we are able to separate the populations by principal component analysis (PCA), demonstrating that ROHs contain information on population structure and the demographic history of a population.

4.4 Study IV: Statistical challenges associated with detecting CNVs using next-generation sequencing (NGS) technology.

4.4.1 Motivation

Whole genome re-sequencing for the identification of CNVs has gained popularity with the recent development of NGS platforms that allow massive parallel sequencing. These techniques have the potential to discover smaller CNVs that were not previously discovered and detect balanced rearrangements such as inversions and translocations. However, analysing NGS data for CNVs is a new and challenging field, with no standard protocols or quality control measures. Also, due to the complexity of the genome and the short read lengths from NGS technology, there are still many challenges associated with the analysis of NGS data for CNVs, no matter which method or algorithm is used.

4.4.2 Results

We describe and discuss areas of potential biases in CNV detection using NGS data, focusing on issues pertaining to (1) mappability, (2) GC-content bias, (3) quality-control measures of reads, and (4) difficulties in identifying duplications. To gain insights to some of the issues discussed, we download real data from the 1000 Genomes Project and analyse its depth of coverage (DOC) data. We show examples of how reads in repeated regions can affect CNV detection, demonstrate current GC correction algorithms, investigate sensitivity of DOC algorithm before and after quality-control of reads and discuss reasons for which duplications are harder to detect than deletions.

Chapter 5 - DISCUSSION

5.1 What makes a good CNV detection method?

The quality of a CNV detection method (including the technology and algorithm) can be broadly attributed to three aspects: (1) location (2) breakpoints (3) genotype. The location and breakpoints of a CNV are closely related, where the breakpoints are given by the start and end positions of a CNV, and the location is the entire region that spans from the start position to the end position. Most studies use the location and breakpoints to determine sensitivity and specificity of a method. However, with SNP/aCGH arrays, the start and end positions are technically not the true start/end positions of a CNV, but rather the start and end probes of the array that was used. Hence, breakpoint precision is highly affected by the resolution of the array. An array with denser probes at and near the location of the CNV will be able to detect the start/end of the CNV with higher precision.

Another less-frequently used criteria for evaluating CNV detection methods is the ability to discern the actual copy number of the region, for example 0 copy versus 1 copy for deletions and 3 or more copies for duplications. This is also known as 'genotyping' of the CNV. Many algorithms use a clustering procedure, assuming that most individuals have normal '2 copies'.

5.2 Concordances among CNV detection methods

From experience of several peer reviews we got during our submission of the manuscripts, many reviewers are often concerned about the low concordance between

the CNVs generated by our methods as compared to the reference list we use. However, this low concordance is often not a very good indicator of bad algorithm performance per se, but rather a more general problem in CNV detection. For example, in McCarroll *et al.* (2008)'s study, they employed a set of very strict criteria on duplicate experiments in SNP arrays to define common CNV regions in eight HapMap samples. Despite that, (on average) 76% of the regions do not overlap with the list of regions found using sequencing. Even when applied to the same raw data, Pinto *et al.* (2011) found that different analytic tools typically yield CNV calls with <50% concordance. The low concordance can be attributed to several factors such as (1) lack of a true gold standard, (2) noisy data resulting in many false identifications and (3) imprecision of the breakpoints identified.

Indeed, the first step of determining the sensitivity of a method is to obtain a 'true positive' dataset. Hence, the first problem with CNV analysis: we do not know the 'true positives'! The closest bet is to use published results from studies that are well-validated as a reference panel, and that is often only possible if you have the same samples as that in the reference panel. HapMap samples are commonly used in methodology research, usually for two main reasons: the raw data are readily available and there are several studies which have characterized the CNV profiles for these individuals and often used as the 'gold standard' (Kidd *et al*., 2008; McCarroll *et al*., 2008; Conrad *et al.*, 2010). When this is not possible, simulation is another way to estimate the sensitivity of the method.

After we have chosen our 'gold standard' dataset, the second difficulty in accessing sensitivity is in answering the question "Is CNV1 and CNV2 the same variant?" In

Figure 5.1a, when the breakpoints of the two variants match perfectly, there is no doubt in calling them the same variant. In Figure 5.1b, the breakpoints are different but the two variant have a good amount of overlap and are of roughly the same length. What about in Figure 5.1c where one breakpoint coincides but the length of the variant differs by a lot? Some studies use a relaxed criterion of calling two variants the same as long as there is a single base overlap, while other studies may be as stringent as requiring at least an 80% reciprocal overlap. A 50% reciprocal overlap seems to be adopted by the majority of studies in recent years. To avoid the need to choose this arbitral percentage, some studies define sensitivity as the proportion of bases that overlap.

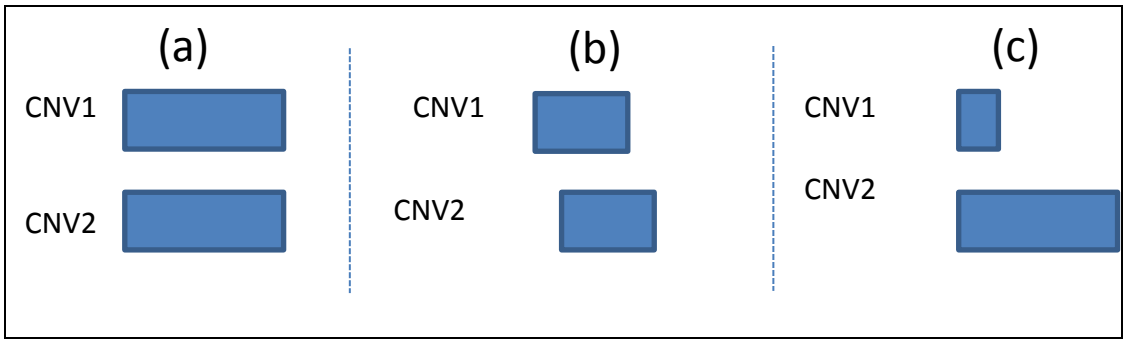

Figure 5.1: Diagram illustrating the non-triviality of determining if two CNVs are the 'same' variant. In (a), CNV1 and CNV2 overlap completely. In this case, we are confident that the two CNVs are the same. In (b), the start and end positions of CNV1 and CNV2 differs, but there is substantial overlap between the two. In (c), CNV1 is completely within the range of CNV2 but the two CNVs differ vastly in lengths. In most research papers, scientists are comfortable with using a 50% reciprocal overlap to determine if two CNVs are concordant.

5.3 Problems caused by repetitive DNA

Repetitive DNA poses challenges in CNV detection regardless whether SNP arrays or sequencing methods are used. For SNP arrays, the density of SNP probes in segmental duplicated regions is sparse due to technical difficulties in assay design and

implementation (Winchester *et al.*, 2009) resulting in a bias against detecting CNVs in segmental duplicated regions using SNP arrays.

For sequencing, reads that fall in repetitive DNA cause problems in alignment and assembly algorithms (Treangen *et al.*, 2012). This problem is exacerbated in NGS (as compared to Sanger sequencing) because the sequenced reads from NGS are relatively short (35-150bp). Furthermore, mutations or sequencing errors in one or two locations may also cause reads to be mapped wrongly (Li *et al.*, 2008). In the 1000 Genomes trios Project, about 20% of the reference genome was considered inaccessible (defined as regions with many ambiguously placed reads or unexpectedly high or low numbers of aligned reads). The resulting low sensitivity in detecting CNVs in repeated/segmental-duplicated regions is a serious problem, because there is an observed enrichment of CNVs in segmental duplicated regions and many breakpoints lie in duplicated regions (Medvedev *et al.*, 2009).

For assembly-based methods, repeat regions create challenges because if the read length is shorter than the repeat region, it is not straightforward to decipher the original sequence since overlap between the reads or contigs will be ambiguous (Knudsen *et al.*, 2010). For other methods that require mapping to a reference, there are different alignment strategies for dealing with multi-reads, such as (1) discarding the reads, (2) choosing a position at random out of all equally good match positions, and (3) reporting all possible positions. In Study IV, we have shown why these strategies are inadequate for dealing with multi-reads.

Recently, there are several algorithms that claim to be able to resolve specific types of CNVs in repeat regions. For example, He *et al*. (2011) developed an algorithm for tandem copy number variation reconstruction in repeat-rich regions, which considers all locations of possible mappings and uses information on read-pair and DOC. Alkan *et al.* (2009) developed a new alignment method, mrFAST. The aligner maps short sequence reads to a repeat-masked reference genome, meaning that all loci with known high-copy common repeats were first masked before alignment, and reports all mapping locations for multi-reads. It also keeps track of mutation in multi-reads. This method has been shown to be able to predict absolute copy number and multicopy differences. Sudmant *et al*. (2010) also uses a similar approach to identify and genotype CNVs within segmental duplications. However, these approaches seem to work only for deeply sequenced data (>20X), and more has to be done to extend these methods for lower coverage data (Chiang *et al*, 2009).

Longer read lengths from third generation sequencing may partially solve the problems with repeats, but even with a read length of 1kb, there still remains about 1.5% of the human genome sequence that is non-unique (Schatz *et al.*, 2010).

5.4 A peek into third generation sequencing (TGS)

Third generation sequencing (TGS) or also known as single molecule sequencing (SMS) promises to improve sequencing rates, throughput and read lengths as compared to NGS. Since it does not require repeated stepwise 'washing and scanning' procedures like in NGS, TGS may increase the sequencing cycle by four orders of magnitude (Eid *et al*., 2009). The first commercially available SMS instrument is the

HeliScope Single Molecular Sequencer by Helicos Biosciences; however, the read lengths are still short at ~32 bases long (Schadt *et al*., 2010). Since PCR amplification is not required in TGS, bias observed in NGS in depth of coverage due to PCR may be resolved. The longer read lengths of TGS will also improve challenges caused by the short read lengths of NGS. Time will reveal if TGS can fulfil its promises for advancement over NGS.

CHAPTER 6 - CONCLUSIONS

- Copy number variations, ROHs and other structural variations are an important source of variation in the human genome, and have been associated with many complex diseases.
- Due to the multi-base and multi-allelic nature of these variants, detecting them with high sensitivity and specificity is still a challenge. Hence, new statistical methods and user-friendly bioinformatics tools are needed for the analyses of these variants.
- In Study I, we develop a method that allows users to detect common CNV regions.
- In Study II, we develop a method that allows users to detect CNVs using information from multiple platforms simultaneously.
- There is a lack of studies investigating regions of homozygosity in Asian populations. There is also a lack of understanding of the relationships between ROHs and haplotype frequency, linkage disequilibrium and positive selection. These are addressed in Study III.
- Next-generation sequencing has the potential to detect CNVs beyond the resolution of SNP arrays and aCGH, as well as detect copy neutral SVs such as inversions and translocations.
- Analytical methods and algorithms for CNV detection using NGS are not yet mature and there are still many challenges. In Study IV, we describe and discuss challenges faced in CNV detection using NGS data.

Chapter 7 – FUTURE DIRECTIONS AND PERSPECTIVES

The field of genetics and genomics has progressed a long way since the first human genome was sequenced in 2000. By now, there are thousands of genes and loci discovered that are associated with simple and complex human diseases, and many of the discoveries were made via GWAS of SNPs. SVs, on the other hand, were much less considered in association studies, particularly attributed to technical difficulties in characterizing SVs with high resolution. Recent development of high-throughput sequencing presents new opportunities for identifying SVs, especially the smaller CNVs that were beyond the resolution of old techniques, as well as copy-neutral events such as inversions and translocations. However, there are still many problems associated with identifying SVs using NGS technology, as discussed in Study IV. As the technology and analytical methods continue to improve, some of these problems may resolve. However, it is of my personal opinion that the following cannot be neglected:

1. Collaborations among various research centres. Even as the cost for whole genome high-throughput sequencing continues to drop, routine sequencing of a large number of individuals will still remain too pricy for the majority of research centres. Collaborations will push the research at a faster pace, overcoming cost and manpower issues. Take for example the 1000 Genomes Project (www.1000genomes.org), which aims to sequence 2500 individuals, and have thus far completed the sequencing of more than 1000 individuals. Such an effort was the result of collaborations of more than 70 research groups and would definitely not have been possible by a single research centre.

2. Well-studied and standardized analysis pipelines and quality-control (QC) metrics. One of the major difficulties in comparing SVs among different studies is that all studies use different algorithms and QC metrics. With NGS technology, there are already numerous algorithms to choose from, but yet no consensus on the appropriate analysis pipeline.
3. Educating a whole new discipline of 'big data biology'. As more and more genomics data are collected, the growing need for storage, processing and analysis of the data becomes more and more apparent. Already, there is a great demand for information technology infrastructure and bioinformatics team to analyse the massive amount of data, with speculations that the costs associated with down-handling, storing and analysis of the data could be more than the production of the data. Hence, we need to train new scientists to handle these upcoming challenges.
4. Beyond discovery studies. Many early works on population wide SVs are 'discovery' studies where SVs in a population are characterized. As our understanding of SVs continues to increase, we should look beyond 'discovery', but aim to collect phenotype data for association studies.
5. Integrated knowledge with RNAseq, transcriptome, proteomics etc. We still do not have a good understanding of the function of SVs in the context of human phenotypes. The integrated knowledge of SVs with transcriptome and proteomics will enhance our ability to interpret the genome.

ACKNOWLEDGEMENTS

Needless to say, the first thanks go out to my supervisors **Chia Kee Seng**, **Yudi Pawitan** and **Agus Salim**, without which this thesis would not have been possible. I sincerely thank them for their patient supervision and support throughout these 4 years and for seeing opportunities in every difficulty we faced. Thank you also for being very flexible supervisors, for giving me the opportunity and privilege to pursue a joint degree as well as attend numerous courses and conferences overseas.

Thank you to my co-authors, **Stefano Calza**, **Ku Chee Seng**, **Vikrant Kumar**, **Anbupalam Thalamuthu**, **Mark Seielstad** and **Nasheen Naidoo**, without which the publications would not have been possible. Special thanks to **Ku Chee Seng** for always patiently answering my numerous questions and for being a fantastic 'walking encyclopedia' on genotyping technologies.

To my mentor **Marie Reilly**, thank you for your care and concern, be it with regards to my academic work or my general well-being, and (together with Yudi), for all the lovely invites to your house and all those yummy dinner treats!

To my undergraduate thesis advisor **Yap Von Bing** who first introduced me to the world of genetics and genomics.

Having been fortunate to pursue my PhD both in Singapore and in Sweden, I would like to give my gratitude to friends and colleagues from the Saw Swee Hock School of Public Health at NUS, as well as from the Department of Medical Epidemiology and Biostatistics at KI.

Special mentions from NUS include **Yang Qian, Suo Chen, Teo Yik Ying, Tai Bee Choo, Sim Xue Ling, Lim Gek Hsiang, Sharon Wee, Kaavya Narasimhalu**, **Tan Chuen Seng, Gao He, Moira Khaw, Katherine Kasiman** and **Salome Rebello**.

From KI, I would like to especially thank **Li Jingmei, Hatef Darabi, Andrea Ganna, Emil Rehnberg, Myeongjee Lee, Tong Gong** and **Ting Zhuang.** To everyone else in MEB, thank you for making MEB such a delightful place to work in!

To my amazing parents and brother, thank you for always believing in me and supporting me in whatever I choose to do.

To all my wonderful friends, especially from RGS symphonic band, RJC ODAC and NUS climbing, you know who you are, thank you for all the fun times!

Last but not least, I would like to acknowledge financial support from the National University of Singapore Graduate School of Science and Engineering (NGSS) scholarship.

REFERENCES

1. Abyzov A *et al.* (2011) CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research* **21**: 974-984.
2. Alkan C *et al.* (2011) Genome structural variation discovery and genotyping. *Nature Review Genetics* **12**: 363-376.
3. Alkan C *et al.* (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genetics* **41**:1061-1067.
4. Aquilante CL *et al.* (2006). Influence of coagulation factor, vitamin K epoxide reductase complex subunit 1, and cytochrome P450 2C9 gene polymorphisms on warfarin dose requirements. *Clinical Pharmacology & Therapeutics* **79**: 291–302.
5. Beroukhim R *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proceedings of the National Academy of Sciences USA* **104**: 20007-20012.
6. Bersaglieri T *et al.* (2004). Genetic Signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics* **74**: 1111-1120.
7. Chen K *et al.* (2009) BreakDancer: An algorithm for high resolution mapping of genomic structural variation. *Naure Methods* **6**: 677–681.
8. Chiang DY, McCarroll SA (2009) Mapping duplicated sequences. *Nature Biotechnology* **27**: 1001-1002.
9. Conrad DF *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature* **464**: 704-712.
10. Curtis D *et al.* (2008) Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. *Annals of Human Genetics* **72**: 261–278.
11. Dalca AV, Brudno M (2010) Genome variation discovery with high-throughput sequencing data. *Briefings in Bioinformatics* **11**: 3-14.

12. Diskin SJ *et al.* (2006) STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Research* **16**:1149-1158.

13. Eid J *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133-138.

14. Gibson J *et al.* (2006) Extended tracts of homozygosity in outbred human populations. *Human Molecular Genetics* **15**, 789-795.

15. Handsaker RE *et al.* (2011) Discovery and genotyping of genome structural polymorphism by sequencing on population scale. *Nature Genetics* **43**: 269-276.

16. Harrington DJ *et al.* (2005) Pharmacodynamic resistance to warfarin associated with a Val66Met substitution in vitamin K epoxide reductase complex subunit 1. *Thrombosis and Haemostasis* **93**: 23-26.

17. Hastings PJ *et al.* (2009) Mechinsms of change in gene copy number. *Nature Review Genetics* **10**: 551-564.

18. He D *et al.* (2011) Efficient algorithms for tandem copy number variation reconstruction in repeat-rich regions. *Bioinformatics* **27**: 1513-1520.

19. Hert DG *et al.* (2008) Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis* **29**: 4618-4626.

20. Huang J *et al.* (2007) Robust smooth segmentation approach for array CGH data analysis. *Bioinformatics* **23:** 2463–2469.

21. Iafrate AJ *et al.* (2004) Detection of large-scale variation in the human genome. *Nature Genetics* **36**: 949-951.

22. Kidd JM *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**:56-64.

23. Knudsen B *et al.* (2010) A computer simulator for assessing different challenges and strategies of de novo sequence assembly. *Genes* **1**: 263-282.

24. Ku CS *et al.* (2010) The discovery of human genetic variations and their use as disease markers: past, present and future. *Journal of Human Genetics* **55**: 403-415.

25. Ku CS *et al.* (2011) Regions of homozygosity and their impact on complex diseases and traits. *Human Genetics* **129**:1-15.

26. Lencz T *et al.* (2007) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proceedings of the National Academy of Sciences USA* **104**: 19942-19947.

27. Li LH *et al.* (2006) Long contiguous stretches of homozygosity in the human genome. *Human Mutation* **27**: 1115-1121.

28. Li H *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* **18**:1851-1858.

29. Locke DP *et al.* (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *American Journal of Human Genetics* **79**: 275-290.

30. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genetics* **24**:133–141.

31. McCarroll SA *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics* **40**: 1166-1174.

32. McQuillan R *et al.* (2008) Runs of homozygosity in European populations. *American Journal of Human Genetics* **83**: 359-372.

33. Medvedev P *et al.* (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods* **6**: S13-20.

34. Medvedev P *et al.* (2010) Detecting copy number variation with mated short reads. *Genome Research* **20**: 1613-1622.

35. Metzker ML (2010) Sequencing technologies – the next generation. *Nature Reviews* **11**:31-46.

36. Mills RE *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* **470**: 59-65.

37. Nalls MA *et al.* (2009) Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. *Neurogenetics* **10**: 183-190.

38. Nothnagel M *et al.* (2010) Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. *Human Molecular Genetics* **19**: 2927-2935.

39. O'Dushlaine CT *et al.* (2010) Population structure and genome-wide patterns of variation in Ireland and Britain. *European Journal of Human Genetics* **18**: 1248-1254.

40. Olshen AB *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**: 557–572.

41. Pang AW *et al.* (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biology* **11**:R52.

42. Pinto *et al.* (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotechnology* **29**: 512-520.

43. Pop M *et al.* (2004) Comparative genome assembly. *Briefings in Bioinformatics* **5**: 237-248.

44. Sabeti PC *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**.

45. Sanger F *et al.* (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences USA* **74**: 5463–5467.

46. Schadt EE *et al.* (2010) A window into third-generation sequencing. *Human Molecular Genetics* **19**: R227-240.

47. Schatz MC *et al.* (2010) Assembly of large genomes using second generation sequencing. *Genome Research* **20**: 1165-1173.

48. Scherer SW *et al.* (2007) Challenges and standards in integrating surveys of structural variation. *Nature Genetics* **39**: S7-15.

49. Schloss JA (2008) How to get genomes at one ten-thousandth the cost. *Nature Biotechnology* **26**: 1113-1115.

50. Sebat J *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science* **305**: 525-528.

51. Sudmant PH *et al.* (2010) Diversity of human copy number variation and multicopy genes. *Science* **330**: 641 – 646.

52. Teo SM *et al.* (2010) Identification of recurrent regions of copy number variation across multiple individuals. *BMC Bioinformatics* **11**:147.

53. Teo SM *et al.* (2011) Multi-platform Segmentation for joint detection of copy number variants. *Bioinformatics* **27**:11.

54. Teo SM *et al.* (2012) Regions of homozygosity in three Southeast Asian populations. *Journal of Human Genetics* **57**: 101-108.

55. Teo YY *et al.* (2009) Genome-wide comparisons of variation in linkage disequilibrium. *Genome Research* **19**: 1849-1860.

56. Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Review Genetics* **13**:36-46.

57. Venter JC *et al.* (2001) The sequence of the human genome. *Science* **291**: 1304 – 1351.

58. Voight BF *et al.* (2006) A map of positive selection in the human genome. *PLoS Biol*ogy **4**: e72.

59. Wain LV *et al.* (2009) Genomic copy number variation, human health, and disease. *Lancet* **374**: 340-350.

60. Wang K *et al.* (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17**:1665-167.

61. Wellcome Trust Case-Control Consortium (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**:713-720.

62. Winchester L *et al.* (2009) Comparing CNV detection methods for SNP arrays. *Briefings in Functional Genomics and Proteomics* **8**: 353-366.

63. Yang TL *et al.* (2010) Runs of homozygosity identify a recessive locus 12q21.31 for human adult height. *The Journal of Clinical Endocrinology & Metabolism* **95**: 3777-3782.

64. Ye K *et al.* (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**: 2865-2871.

65. Yoon S *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research* **19**:1586-1592.

66. Zhang NR *et al.* (2010) Joint estimation of DNAcopy number from multiple platforms. *Bioinformatics* **26**: 153-160.

67. Zhu Y *et al*. (2007) Estimation of Wafarin Maintenance Dose Based on VKORC1 (-1639 G>A) and CYP2C9 Genotypes. *Clinical Chemistry* **53**: 1199-1205.